TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Bài báo khoa học

Xây dựng mô hình dự báo BOD_5 cho hạ lưu sông Sài Gòn – Đồng Nai dựa trên các mạng nơ–ron nhân tạo MLP và RBF

Nguyễn Thị Diễm Thúy[1*], Phạm Thị Thảo Nhi[2], Đoàn Thị Trúc Măn[3], Đào Nguyên Khôi [4]

[1] Viện Môi trường và Tài nguyên, Đại học Quốc gia TP.HCM; nguyenthidiemthuyapag@gmail.com;
[2] Viện khoa học và Công nghệ tính toán Tp.HCM; ptthaonhi@gmail.com
[3] Đài khí tượng thủy văn khu vực Nam bộ; trucmandoan@gmail.com
[4] Khoa Môi trường, Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia TP.HCM; dnkhoi@hcmus.edu.vn
***** Tác giả liên hệ: nguyenthidiemthuyapag@gmail.com; Tel.: +84–968638978

Ban Biên tập nhận bài: 11/3/2020; Ngày phản biện xong: 3/4/2021; Ngày đăng bài: 25/4/2021

Tóm tắt: Nhu cầu oxy sinh hóa (BOD) là thông số chất lượng nước quan trọng để đánh giá mức độ ô nhiễm của nước trên các sông, hồ. Tuy nhiên, việc xác định nồng độ BOD_5 trong nước theo các phương pháp phân tích trong phòng thí nghiệm thường mất nhiều thời gian (5 ngày). Mục tiêu của nghiên cứu là xây dựng mô hình dự báo thông số BOD_5 dựa trên hai mô hình nơ–ron nhân tạo là MLP và RBF tại hạ lưu sông Sài Gòn–Đồng Nai và đánh giá hiệu quả dự báo giữa hai mô hình. Bảy kịch bản được xây dựng dựa trên tương quan riêng phần giữa thông số BOD_5 với các thông số chất lượng nước khác bao gồm COD, DO, TSS, Coliform, $P–PO_4^{3–}$, T và $N–NH_4^+$. Bộ dữ liệu bao gồm 08 thông số chất lượng nước theo tháng từ 2013–2018 và được chia thành hai phần theo tỷ lệ 75:25 phục vụ huấn luyện và kiểm tra các mô hình. Kết quả nghiên cứu cho thấy, cả hai mô hình MLP và RBF đều có khả năng dự báo tốt BOD_5 tại khu vực, tuy nhiên mô hình RBF với 05 thông số đầu vào (COD, DO, TSS, Coliform, $P–PO_4^{3–}$) cho kết quả dự báo tốt nhất với NSE = 0,855, R^2 = 0,9, RMSE = 0,529 cho quá trình huấn luyện và NSE = 0,848, R^2 = 0,865, RMSE = 0,454 cho quá trình kiểm tra. Kết quả nghiên cứu này cũng là nền tảng phục vụ cho việc dự báo các thông số chất lượng nước khác, cũng như dự báo ngắn hạn BOD_5 trong khu vực nghiên cứu.

Từ khóa: Nhu cầu oxy sinh hóa; Mô hình nơ–ron nhân tạo; MLP; RBF; Hạ lưu sông Sài Gòn–Đồng Nai

1. Mở đầu

Nước là nguồn tài nguyên quan trọng, thiết yếu trong cuộc sống con người và sự phát triển của đất nước. Chất lượng nước là một chỉ tiêu quan trọng liên quan đến tất cả khía cạnh của hệ sinh thái và đời sống con người, như sức khỏe cộng đồng, sản xuất lương thực, hoạt động kinh tế và đa dạng sinh học. Do đó, chất lượng nước cũng là một trong những cơ sở để đánh giá mức độ đói nghèo, thịnh vượng và trình độ văn hoá của khu vực. Trong đó, nhu cầu oxy sinh hóa (BOD) là một trong những thông số chất lượng nước quan trọng, cho phép đánh giá mức độ ô nhiễm hữu cơ có khả năng phân hủy sinh học dưới điều kiện hiếu khí, đây là thông số quan trọng để đánh giá mức độ ô nhiễm của nước, BOD càng cao chứng tỏ lượng chất hữu cơ có khả năng phân hủy sinh học trong nước ô nhiễm càng lớn. Trong thực tế, khó

xác định được toàn bộ lượng oxy cần thiết để các vi sinh vật phân hủy các chất hữu cơ có trong nước mà chỉ xác định được lượng oxy cần thiết trong 5 ngày ở nhiệt độ 20^{o}C trong bóng tối [1].

Các phương pháp đo lường truyền thống thường phụ thuộc vào phân tích trong phòng thí nghiệm, mất nhiều thời gian [2–3]. Cụ thể, đối với thông số BOD_5 mất khoảng 5 ngày để có được giá trị BOD theo các phương pháp đo lường hóa học thông thường. Đối với các công cụ giám sát trực tuyến có thể cho kết quả quan trắc liên tục, tuy nhiên cần chi phí kinh tế cao. Vì vậy, mô hình dự báo và dự báo chất lượng nước là rất cần thiết để theo dõi liên tục các thông số chất lượng nước trên sông, cũng như đóng vai trò rất quan trọng trong công tác quản lý tài nguyên nước. Hiện nay có nhiều phương pháp khác nhau để mô hình hóa và dự đoán chất lượng nước như mô hình khái niệm, mô hình vật lý, mô hình số, mô hình thống kê, v.v.; tuy nhiên trong những năm gần đây, mô hình trí tuệ nhân tạo (AI) đã được sử dụng vì tính đơn giản và tính chính xác của kết quả dự báo. Một điểm mạnh nữa của mô hình AI là mô hình AI có khả năng dự báo các hiện tượng phức tạp và phi tuyến tính mà không cần hiểu rõ về bản chất vấn đề. Do đó, việc sử dụng phương pháp tiếp cận AI trong dự báo chất lượng nước trở thành một hướng nghiên cứu tiềm năng và thu hút sự quan tâm của nhiều nhà nghiên cứu trên thế giới.

Một số nghiên cứu điển hình có thể kể đến như nghiên cứu của Dogan và cộng sự năm 2008 đã sử dụng mô hình nơ–ron nhân tạo (ANN) để dự báo BOD theo ngày, kết quả cho thấy mô hình ANN có khả năng dự báo BOD tốt dựa trên 04 thông số COD, SS, lưu lượng và nitơ với sai số trung bình 10,03% [4]. Nghiên cứu của Csábrági và cộng sự năm 2018 đã dự báo nồng độ DO dựa vào các thông số pH, độ dẫn điện, nhiệt độ và dòng chảy bằng các mạnh nơ – ron nhân tạo bao gồm các mô hình tuyến tính (MLR) và phi tuyến tính (MLP, RBF và GR), kết quả cho thấy các mô hình phi tuyến tính có khả năng dự báo DO tốt hơn so với mô hình tuyến tính và mô hình RBF có hiệu quả dự báo tốt nhất trong tất cả các mô hình với chỉ số RMSE = 1,63 và R^2 = 0,59 [5]. Một số nghiên cứu khác như nghiên cứu của Dara và cộng sự năm 2018 đã sử dụng mô hình MLP với 10 thông số chất lượng nước đầu vào, 1 lớp ẩn – 5 nodes và 1 lớp đầu ra để dự báo BOD [6], nghiên cứu dự báo các thông số chất lượng nước (TSS và BOD) bằng mô hình hồi quy tuyến tính và mô hình mạng nơ–ron nhân tạo (Deep Feedforward Network) của Ahamad và cộng sự năm 2019 [7].

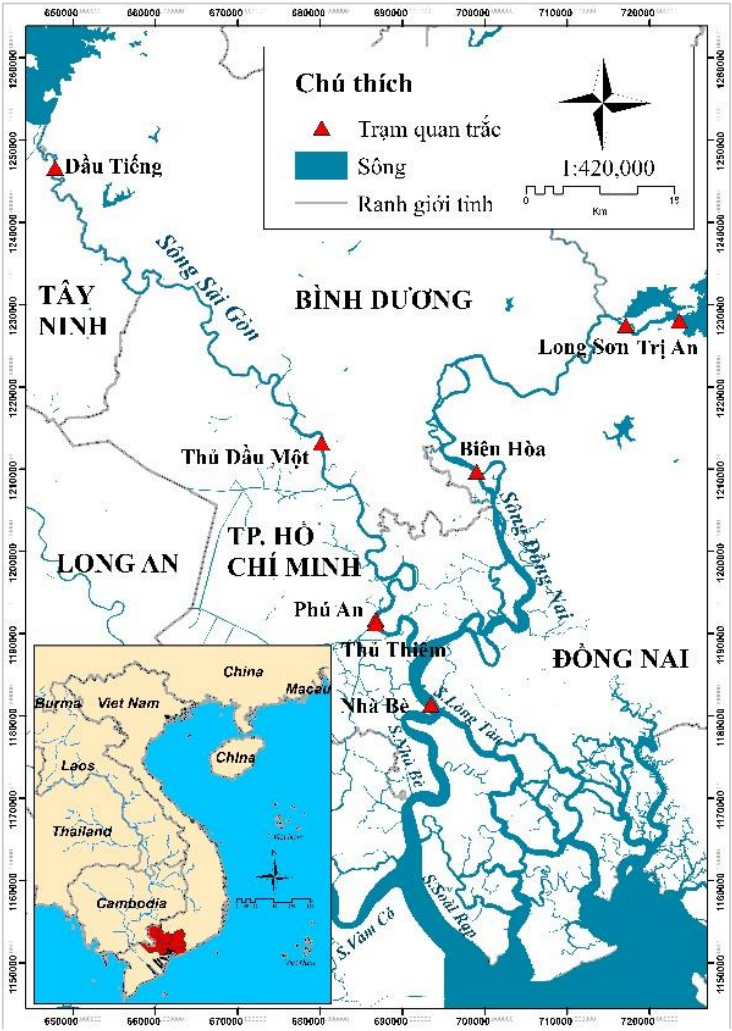
Từ các nghiên cứu đã thực hiện có thể thấy, mạng nơ–ron nhân tạo có khả năng dự báo và dự báo tốt các thông số chất lượng nước trên sông vì vậy trong nghiên cứu này hai mạng nơ–ron nhân tạo là MLP (*Multilayer perceptrons*) và RBF (*Radial basic function*) được sử dụng để dự báo thông số BOD_5 tại hạ lưu sông Sài Gòn–Đồng Nai, đây là khu vực chịu nhiều ảnh hưởng của hoạt động phát triển công nghiệp và đô thị của vùng kinh tế trọng điểm phía nam. Mục tiêu của nghiên cứu là dự báo nồng độ BOD_5 tại hạ lưu sông Sài Gòn–Đồng Nai dựa trên mạng MLP và RBF và so sánh hiệu quả dự báo giữa hai mô hình. Để đạt được mục tiêu trên các nội dung cơ bản được thực hiện để xây dựng một mô hình nơ–ron nhân tạo trong nghiên cứu này bao gồm: (1) thu thập và tiền xử lý dữ liệu; (2) lựa chọn đầu vào, (3) xử lý và phân tách dữ liệu, (4) lựa chọn kiến trúc mô hình, (5) huấn luyện mô hình và (6) kiểm định mô hình để tìm ra bộ thông số tối ưu của các mô hình [8].

2. Phương pháp nghiên cứu

2.1 Khu vực nghiên cứu

Khu vực nghiên cứu thuộc vùng hạ lưu sông Sài Gòn–Đồng Nai, nằm ở kinh độ 10°30’ – 11°30’ B và vĩ độ 106°15’ – 107°15’ Đ (Hình 1). Khu vực nghiên cứu có diện tích khoảng 3.200 km^2 đi qua các tỉnh Bình Phước, Bình Dương, Tây Ninh, Long An, Đồng Nai và thành phố Hồ Chí Minh (TP.HCM). Bao gồm các con sông chính như hạ lưu sông Đồng Nai, sông Soài Rạp, sông Nhà Bè, sông Sài Gòn, sông Vàm Cỏ và các sông, kênh thuộc huyện Cần Giờ, TP.HCM.

Khí hậu của khu vực nghiên cứu là nhiệt đới gió mùa, với lượng mưa trung bình năm khá cao, khoảng 1.800 mm. Có hai mùa riêng biệt là mùa mưa (tháng 4 đến tháng 10) và mùa khô (tháng 11 đến tháng 3 năm sau), trong đó lượng mưa trong mùa mưa chiếm khoảng 80–85% tổng lượng mưa năm. Do nằm ở hạ lưu hệ thống sông Sài Gòn–Đồng Nai nên dòng chảy chịu sự chi phối mạnh mẽ bởi thủy triều biển Đông với cơ chế dòng chảy chính là dòng chảy 2 chiều. Bên cạnh đó, hạ lưu sông Sài Gòn–Đồng Nai chảy qua TP.HCM, Đồng Nai, Bình Dương, Bà Rịa–Vũng Tàu, đây được xem như một vùng kinh tế giàu tiềm năng, vùng kinh tế động lực mạnh hàng đầu của Việt Nam hiện nay và trong nhiều năm tới [9]. Dưới ảnh hưởng của hoạt động phát triển công nghiệp và đô thị, vấn đề ô nhiễm nước mặt đã và đang là một trong những vấn đề bức thiết của khu vực này. Vì vậy, khu vực này được chọn làm khu vực nghiên cứu.



Hình 1. Khu vực nghiên cứu.

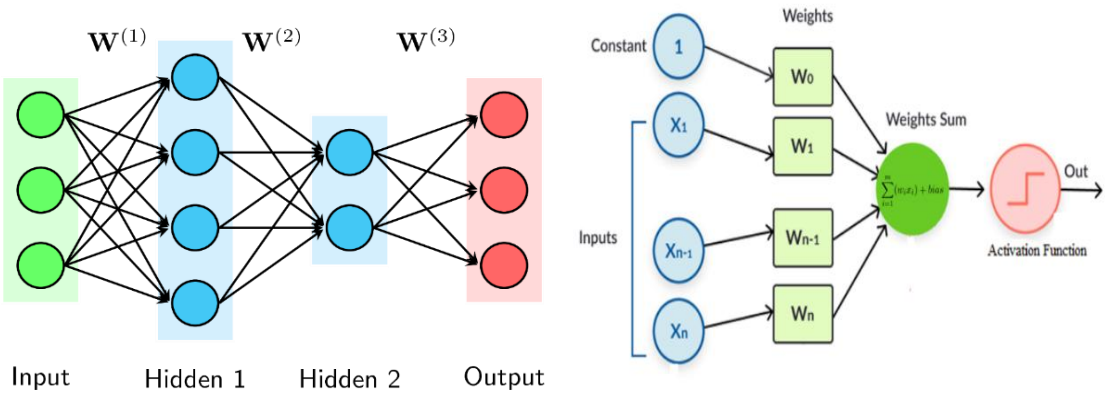
2.2. Phương pháp mô hình hóa

2.2.1. Mô hình Multilayer Perceptrons (MLP)

Mạng nơ–ron nhân tạo là một dạng trí tuệ nhân tạo dựa trên chức năng của bộ não và hệ thần kinh của con người. Một mạng nơ–ron nhân tạo có hai thành phần cơ bản là nơ–ron và liên kết. Một nơ–ron là phần tử xử lý và một liên kết được sử dụng để kết nối một nơ–ron này với một nơ–ron khác, mỗi liên kết có trọng số riêng của nó. Mạng nơ–ron chỉ lan truyền theo hướng thuận từ lớp đầu vào qua một hoặc nhiều lớp ẩn đến lớp đầu ra được gọi là mạng

nơ–ron lan truyền thẳng. Cả hai mô hình MLP và RBF được xây dựng trong nghiên cứu đều là mạng nơ–ron lan truyền thẳng.

Multilayer perceptron (MLP) là mạng nơ–ron lan truyền thẳng phổ biến nhất. Mô hình MLP được sử dụng rộng rãi trong các bài toán dự báo các yếu tố khí tượng, thủy văn và chất lượng nước. MLP bao gồm nhiều lớp tế bào thần kinh (nơ–ron) tương tác với các kết nối có trọng số [10]. Nói chung, một mô hình MLP bao gồm một lớp đầu vào (*input layer*), một hoặc một số lớp ẩn (*hidden layers*) và một lớp đầu ra (*output layer*). Hình 2a thể hiện cấu trúc của mạng MLP với 2 lớp ẩn và Hình 2b trình bày các thành phần chính trong mô hình MLP.

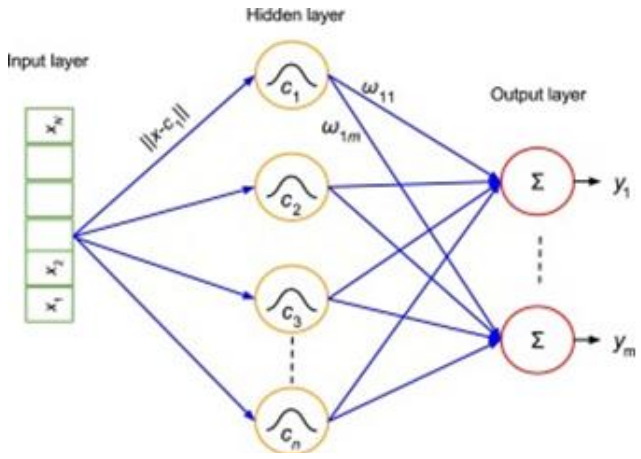


(a) Cấu trúc mô hình (b) Thành phần chính trong MLP

Hình 2. Cấu trúc mạng MLP.

2.2.2. Mô hình Radial Basic Function (RBF)

Radial Basic Function (RBF) là một mạng nơ–ron lan truyền thẳng bao gồm 03 lớp chính: lớp đầu vào, lớp ẩn và lớp đầu ra. Số lượng nơ–ron trong lớp đầu vào phụ thuộc vào chiều của vector đầu vào, số lượng nơ–ron trong lớp đầu ra phụ thuộc vào số nhãn trong dữ liệu. Số lượng nơ–ron trong lớp ẩn quyết định cấu trúc của mạng. Hình 3 thể hiện cấu trúc của mạng RBF. Mô hình RBF có cấu trúc đơn giản và tốc độ học nhanh hơn so với mô hình MLP [11].



Hình 3. Cấu trúc mạng RBF [12].

Quy trình tính toán trong mô hình RBF được thực hiện qua các bước chính sau:

+ Dữ liệu đầu vào được đưa vào mạng thông qua lớp đầu vào.

+ Sau đó mỗi nơ–ron trong lớp ẩn tính toán sự tương đồng giữa dữ liệu đầu vào và nguyên mẫu lưu trữ trong mỗi nơ–ron, càng nhiều kết quả nguyên mẫu kết quả càng chính xác. Mỗi nơ–ron trong lớp ẩn có một hàm kích hoạt Gaussian, với công thức như sau:

$$\phi\left(\left\|x-c_j\right\|\right)=e^{-\left(\frac{\left\|x-c_j\right\|^2}{2\sigma_j^2}\right)} \quad (1)$$

Trong đó x là vector đầu vào; c_j là tâm hàm Gaussian và σj là bề rộng hàm Gaussian của nơ–ron thứ j.

+ Đầu ra của RBF được tính toán sử dụng phương pháp trọng số trung bình theo công thức sau:

$$y_i=\sum_{j=1}^{n} W_{ij}\phi_j\left(x\right) \quad (2)$$

Trong đó W_{ij} là trọng số thứ i giữa lớp ẩn và lớp đầu ra; n là số lượng nơ–ron trong lớp ẩn.

2.3. Thu thập, xử lý và phân chia dữ liệu

2.3.1. Thu thập và chuẩn hóa dữ liệu

Các dữ liệu được sử dụng làm dữ liệu đầu vào cho mô hình dự báo BOD_5 tại khu vực hạ lưu sông Sài Gòn–Đồng Nai bao gồm 08 thông số chất lượng nước nhu cầu oxy sinh hóa (BOD_5), oxy hòa tan (DO), nhu cầu oxy hóa học (COD), nhiệt độ (T), amoni ($N–NH_4^+$), phốt phát ($P–PO_4^{3-}$), tổng chất rắn lơ lửng (TSS) và Tổng coliform (Coliform) tại 08 trạm quan trắc theo tháng từ năm 2013–2018 được thu thập từ Đài Khí tượng Thủy văn khu vực Nam bộ. Vị trí các trạm quan trắc chất lượng nước được thể hiện trong Hình 1 và Bảng 1 thể hiện mô tả thống kê của các dữ liệu chất lượng nước trong khu vực nghiên cứu.

Bảng 1. Đặc trưng các thông số chất lượng nước tại khu vực nghiên cứu.

Thông số	Đơn vị	Lớn nhất	Nhỏ nhất	Trung bình	Độ lệch chuẩn
BOD_5	mg/l	11,00	1,00	3,54	1,34
T	^oC	27,40	25,00	26,07	0,59
DO	mg/l	8,00	1,00	4,35	2,03
TSS	mg/l	482,00	5,60	24,95	41,91
COD	mg/l	26,00	3,00	12,52	4,93
$P–PO_4^{3-}$	mg/l	1,84	0,00	0,02	0,12
$N–NH_4^+$	mg/l	48,10	0,01	0,96	4,30
Coliform	MPN/100ml	24.000,00	230,00	5.730,28	6.824,18

Để thực hiện các phép tính trong mô hình (cộng, nhân ma trận, vector) yêu cầu dữ liệu đầu vào có cùng kích thước, vì vậy việc chuẩn hóa dữ liệu đầu vào bao gồm loại bỏ các dữ liệu nhiễu và chuẩn hóa các dữ liệu về cùng khoảng giá trị là việc làm quan trọng trước khi thực hiện dự báo BOD_5 dựa trên các mạng nơ–ron nhân tạo. Dựa vào đặc điểm của bộ dữ liệu thu thập, nghiên cứu sử dụng phương pháp chuẩn hóa min–max để chuẩn hóa dữ liệu đầu vào của mô hình, đây là phương pháp đơn giản nhằm đưa tất cả các đặc trưng về cùng một khoảng giá trị. Công thức cụ thể như sau:

$$x_i^{'}=\frac{x_i-\min(x_i)}{\max(x_i)-\min(x_i)} \quad (3)$$

Trong đó x_i và $x^{'}_i$ lần lượt là giá trị đặc trưng ban đầu và giá trị đặc trưng sau khi được chuẩn hóa; min (x_i) và max (x_i) là giá trị nhỏ nhất và lớn nhất của đặc trưng thứ i xét trên toàn bộ các điểm dữ liệu của tập dữ liệu.

2.3.2. Phân chia dữ liệu

Một trong những bước quan trọng nhất trong việc phát triển mô hình là quá trình chia nhỏ dữ liệu để huấn luyện và kiểm tra. Dữ liệu huấn luyện sẽ được sử dụng để xác định cấu trúc mô hình, cũng như giá trị của các tham số mô hình và bộ dữ liệu kiểm tra được sử dụng để đánh giá hiệu quả của mô hình.

Bước này thường được thực hiện bằng cách thử và sai để đảm bảo rằng mô hình có thể đạt được hiệu quả tối ưu [13]. Sau khi thực hiện, bộ dữ liệu được chia thành hai phần phục vụ cho quá trình huấn luyện và kiểm tra, cụ thể 75% (185 dữ liệu/1 thông số) được sử dụng cho quá trình huấn luyện và 25% (61 dữ liệu/1 thông số) được sử dụng cho quá trình kiểm tra mô hình, tỷ lệ này cũng đã được áp dụng và đạt hiệu quả cao trong một số nghiên cứu ứng dụng mô hình trí tuệ nhân tạo để dự báo chất lượng nước như nghiên cứu [14–16].

2.4. Đánh giá hiệu quả dự báo của mô hình

Hiệu quả dự báo của các mô hình được đánh giá bằng phương pháp đồ thị và phương pháp thống kê để so sánh chất lượng và độ tin cậy của kết quả dự báo với số liệu thực đo. Trong nghiên cứu này, các phương pháp thống kê đánh giá kết quả mô hình bao gồm hệ số tương quan (R^2), hệ số hiệu quả Nash–Sutcliffe (NSE) và sai số quân phương (RMSE). Giá trị của R^2 và NSE càng gần 1 thì mô hình càng đạt hiệu quả cao, và giá trị RMSE càng gần 0 thì mô hình có sai số càng nhỏ [17].

2.5. Xây dựng các kịch bản dự báo

Các thông số chất lượng nước đã thu thập được sử dụng để xây dựng các kịch bản tính toán dựa trên tương quan riêng phần giữa BOD_5 với các thông số chất lượng nước khác giai đoạn 2013–2018. Bảng 2 thể hiện kết quả tính tương quan riêng phần giữa BOD_5 với 07 thông số chất lượng nước đầu vào còn lại, kết quả cho thấy BOD_5 có tương quan cao nhất với thông số COD (0,85), tiếp đó là thông số DO (–0,55) và thông số amoni có tương quan thấp nhất.

Bảng 2. Tương quan giữa các thông số đầu vào và BOD_5.

Thông số	COD	DO	TSS	Coliform	$P–PO_4^{3–}$	T	$N–NH_4^+$
Tương quan (r)	0,85	–0,55	0,21	0,17	0,07	–0,07	0,01

Các kịch bản với tổ hợp thông số đầu vào được xây dựng dựa trên mức độ tương quan từ cao đến thấp của các thông số đầu vào, theo đó 07 kịch bản dự báo trong nghiên cứu được mô tả như trong Bảng 3.

Bảng 3. Các kịch bản dự báo BOD_5.

STT	**Kịch bản**	**Thông số đầu vào**
1	KB1	COD
2	KB2	COD, DO
3	KB3	COD, DO, TSS
4	KB4	COD, DO, TSS, Coliform
5	KB5	COD, DO, TSS, Coliform, $P–PO_4^{3–}$
6	KB6	COD, DO, TSS, Coliform, $P–PO_4^{3–}$, T
7	KB7	COD, DO, TSS, Coliform, $P–PO_4^{3–}$, T và $N–NH_4^+$

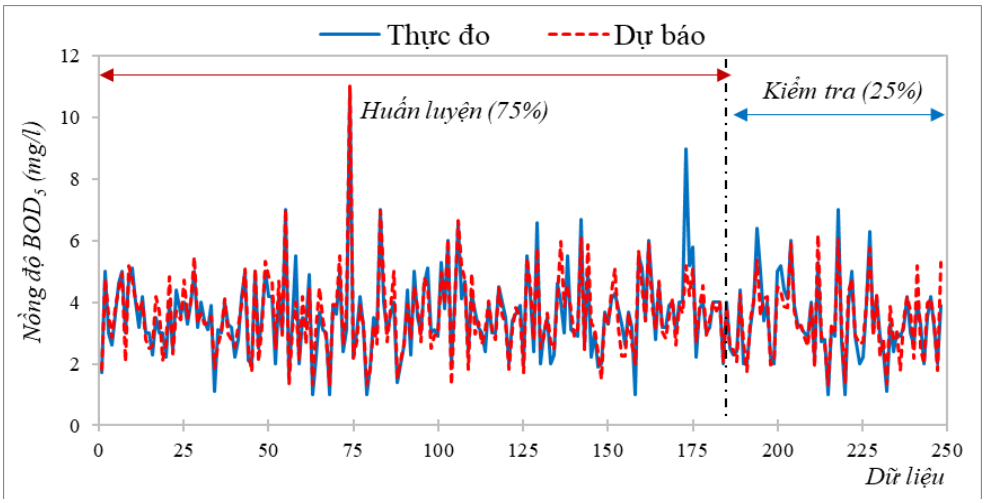
3. Kết quả và thảo luận

3.1. Mô hình MLP

Mô hình MLP được sử dụng để dự báo BOD_5 theo 07 kịch bản với các thông số đầu vào khác nhau, hiệu quả dự báo theo các chỉ số RMSE, NSE và R^2 trong tất cả các kịch bản được thể hiện trong Bảng 4. Kết quả cho thấy, các mô hình MLP với các thông số đầu vào khác nhau đều cho kết quả dự báo BOD_5 khá tốt với RMSE < 0,813, R^2 > 0,740 và NSE > 0,723 cho cả hai giai đoạn huấn luyện và kiểm tra. Trong đó, kịch bản KB7 với 07 thông số đầu vào bao gồm COD, DO, TSS, Coliform, $P–PO_4^{3–}$, T và $N–NH_4^+$ cho kết quả dự báo tốt nhất với NSE, R^2 lớn nhất và RMSE nhỏ nhất so với 06 kịch bản còn lại, cụ thể chỉ số NSE = 0,834, R^2 = 0,836 và RMSE = 0,551 cho quá trình huấn luyện và NSE = 0,832, R^2 = 0,832 và RMSE = 0,521 cho quá trình kiểm tra. Đồ thị so sánh giữa nồng độ BOD_5 thực đo và dự báo trong quá trình huấn luyện và kiểm tra theo KB7 được thể hiện trong Hình 4.

Bảng 4. Hiệu quả dự báo BOD_5 của mô hình MLP.

Kịch bản		KB7	KB6	KB5	KB4	KB3	KB2	KB1
Huấn luyện	RMSE	**0,551**	0,571	0,573	0,584	0,589	0,605	0,655
	R^2	**0,836**	0,814	0,819	0,816	0,801	0,805	0,740
	NSE	**0,834**	0,808	0,818	0,811	0,799	0,802	0,729
Kiểm tra	RMSE	**0,521**	0,632	0,570	0,573	0,647	0,556	0,813
	R^2	**0,832**	0,808	0,813	0,816	0,789	0,810	0,744
	NSE	**0,832**	0,805	0,811	0,808	0,784	0,802	0,723



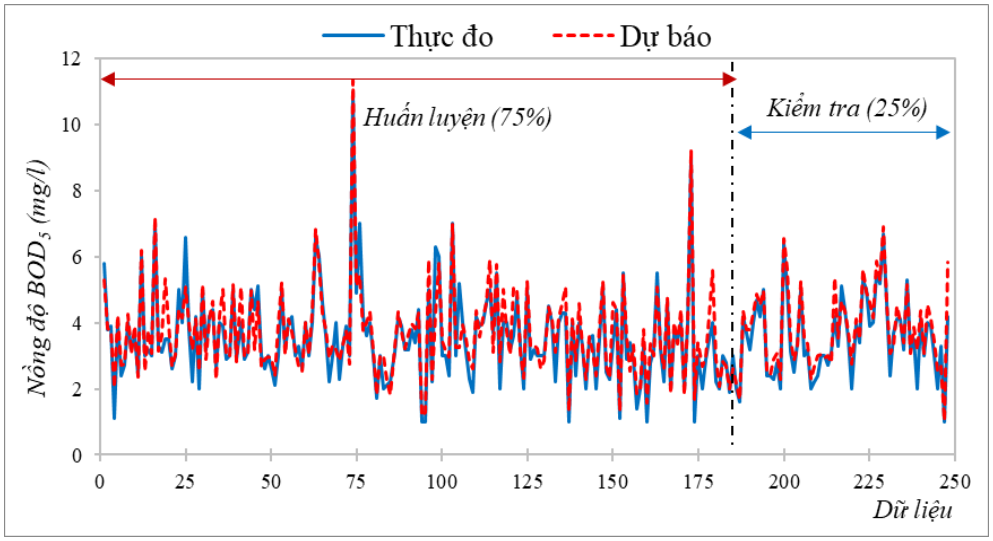
Hình 4. Kết quả dự báo BOD_5 của mô hình MLP–KB7.

3.2. Mô hình RBF

Hiệu quả dự báo BOD_5 theo 07 kịch bản tính toán của mô hình RBF được trình bày trong Bảng 5. Kết quả cho thấy, mô hình RBF theo kịch bản KB5 với 05 thông số đầu vào (COD, DO, TSS, Coliform, $P–PO_4^{3–}$) cho hiệu quả dự báo tốt nhất với các giá trị NSE, R^2 lớn nhất và giá trị RMSE nhỏ nhất trong tất cả các kịch bản ở cả quá trình huấn luyện và kiểm tra, cụ thể chỉ số NSE = 0,855, R^2 = 0,9, RMSE = 0,529 cho quá trình huấn luyện và chỉ số NSE = 0,848, R^2 = 0,865, RMSE = 0,454 cho quá trình kiểm tra. Hình 5 thể hiện kết quả dự báo BOD_5 của mô hình RBF theo KB5.

Bảng 5. Hiệu quả dự báo BOD_5 của mô hình RBF.

Kịch bản		KB7	KB6	KB5	KB4	KB3	KB2	KB1
Huấn luyện	RMSE	0,566	0,492	**0,529**	0,553	0,607	0,590	0,618
	R^2	0,828	0,854	**0,900**	0,839	0,803	0,807	0,758
	NSE	0,827	0,849	**0,855**	0,836	0,802	0,804	0,749
Kiểm tra	RMSE	0,657	0,616	**0,454**	0,493	0,559	0,610	0,803
	R^2	0,735	0,845	**0,865**	0,857	0,798	0,796	0,785
	NSE	0,730	0,836	**0,848**	0,835	0,796	0,792	0,746



Hình 5. Kết quả dự báo BOD_5 của mô hình RBF–KB5.

Bên cạnh đó, kết quả còn cho thấy rằng việc tăng số lượng thông số đầu vào không phải lúc nào cũng cho hiệu quả dự báo tốt hơn, bằng chứng là hiệu quả dự báo của KB5 với 05 thông số dầu vào tốt hơn so với KB6 và KB7 với lần lượt 06 và 07 thông số đầu vào.

3.3. Cấu trúc và bộ thông số tối ưu của mô hình đã xây dựng

Kết quả thống kê hiệu quả dự báo từ hai mô hình trong Bảng 4 và Bảng 5 cho thấy rằng mô hình RBF có khả năng dự báo BOD_5 tại khu vực nghiên cứu tốt hơn so với mô hình MLP, cụ thể hiệu quả dự báo của mô hình RBF tốt hơn thông qua các chỉ số thống kê NSE, R^2 và RMSE. Bên cạnh đó, mô hình RBF chỉ sử dụng 05 thông số (COD, DO, TSS, Coliform, $P–PO_4^{3–}$) để cho hiệu quả dự báo tốt nhất, ngược lại mô hình MLP cần dùng 07 thông số đầu vào (COD, DO, TSS, Coliform, $P–PO_4^{3–}$, T và $N–NH_4^+$) để cho kết quả tốt nhất. Việc giảm số lượng thông số đầu vào có thể tiết kiệm được chi phí phân tích và tăng hiệu quả kinh tế.

Cấu trúc và bộ tham số tối ưu của mô hình RBF theo kịch bản B5 với 05 thông số đầu vào là COD, DO, TSS, Coliform và $P–PO_4^{3–}$ được thể hiện trong Bảng 6. Cụ thể, mô hình RBF với 07 lớp ẩn, số lượng nơ–ron trong các lớp khác nhau, hàm Relu được chọn là hàm kích hoạt với tỷ lệ học = 0,001, Epsolon = 1e–07 và Beta = 1. Thuật toán tối ưu được sử dụng là RMSprop.

4. Kết luận

Nghiên cứu đã thực hiện dự báo BOD_5 tại hạ lưu sông Sài Gòn–Đồng Nai dựa trên hai mạng nơ–ron nhân tạo là MLP và RBF. Bảy kịch bản sử dụng để dự báo BOD_5 tại khu vực nghiên cứu được xây dựng dựa trên tương quan riêng phần giữa thông số BOD_5 với các thông

số chất lượng nước khác bao gồm COD, DO, TSS, Coliform, $P–PO_4^{3–}$, T và $N–NH_4^+$. Hiệu quả dự báo của các mô hình được đánh giá thông qua ba chỉ số NSE, R^2 và RMSE.

Kết quả nghiên cứu cho thấy, cả hai mô hình đều có khả năng dự báo tốt BOD_5 tại khu vực với RMSE < 0,85, R^2>0,74 và NSE>0,72 ở tất cả các kịch bản tính toán, tuy nhiên mô hình RBF có hiệu quả dự báo cao hơn hơn so với mô hình MLP. Ngoài ra, mô hình RBF còn sử dụng ít thông số đầu vào hơn, cụ thể mô hình RBF với 05 thông số đầu vào cho hiệu quả dự báo tối ưu nhất, trong khi mô hình MLP cần 07 thông số đầu vào để cho kết quả dự báo tối ưu. Cấu trúc và bộ tham số tối ưu của mô hình tìm được sau quá trình huấn luyện và kiểm tra trong dự báo hiện trạng BOD_5 cũng là nền tảng phục vụ cho việc dự báo các thông số chất lượng nước khác cũng như dự báo ngắn hạn các thông số chất lượng nước trong tương lai.

Các kết quả đạt được trong nghiên cứu cho thấy thế mạnh của các thuật toán trí tuệ nhân tạo trong dự báo các thông số chất lượng nước. Tuy nhiên, trong các nghiên cứu tiếp theo sẽ tiến hành thử nghiệm các kỹ thuật AI khác để tìm ra kỹ thuật tốt nhất nhằm tăng độ chính xác của kết quả dự báo.

Bảng 6. Cấu trúc và tham số tối ưu mô hình RBF dự báo BOD_5.

	Số lớp ẩn	Loại lớp ẩn	Số nơ–ron		Tham số	Giá trị
Cấu trúc mô hình	Lớp ẩn 1	RBFLayer	35	**Tham số mô hình**	Hàm kích hoạt	Relu
	Lớp ẩn 2	Dense	25		Tỷ lệ học	0,001
	Lớp ẩn 3	Dense	25		Epsilon	1,00e–07
	Lớp ẩn 4	Dense	15		Beta	1
	Lớp ẩn 5	Dense	5		Thuật toán tối ưu	RMSprop
	Lớp ẩn 6	Dense	5		Epoch	500
	Lớp ẩn 7	Dense	5		Batch–size	50

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: N.T.D.T., D.N.K.; Lựa chọn phương pháp nghiên cứu: N.T.D.T., D.N.K.; Xử lý số liệu: P.T.T.N., D.T.T.M.; Viết bản thảo bài báo: N.T.D.T., D.N.K.; Chỉnh sửa bài báo: N.T.D.T., D.N.K., P.T.T.N., D.T.T.M.

Lời cảm ơn: Nghiên cứu này được thực hiện dưới sự tài trợ của Sở Khoa Học và Công Nghệ Tp.HCM và được thực hiện bởi Viện Khoa học và Công nghệ Tính toán (ICST) thông qua Hợp đồng thực hiện nhiệm vụ khoa học và công nghệ số 11/2020/HĐ–QPTKHCN ngày 22 tháng 04 năm 2020.

Lời cam đoan: Tập thể tác giả cam đoan bài báo này là công trình nghiên cứu của tập thể tác giả, chưa được công bố ở đâu, không được sao chép từ những nghiên cứu trước đây; không có sự tranh chấp lợi ích trong nhóm tác giả.

Tài liệu tham khảo

1. Đặng Kim Chi. Hoá học môi trường. NXB KHKT, Hà Nội. 1999.
2. Han, H.G; Qiao, J.F. A self–organizing fuzzy neural network based on a growing–and–pruning algorithm. *IEEE Trans. Fuzzy Syst.* **2010**, *18*, 1129–1143.
3. Han, H.G.; Qiao, J.F. A structure optimisation algorithm for feedforward neural network construction. *Neurocomputing* **2013**, *99*, 347–357.
4. Dogan, E.; Ates, A.; Yilmaz, C.; Eren, B. Application of Artificial Neural Networks to Estimate Wastewater Treatment Plant Inlet Biochemical Oxygen Demand. *Environ. Prog.* **2008**, *27*, 439–446.
5. Csábrági, A.; Molnár, S.; Tanos, P.; Kovács, J. Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube. *Ecol. Eng.* **2017**, *100*, 63–72.

6. Dara, F.; Devolli, A.; Kodra, A. An artificial neural networks modell for predicting BOD of Ishëm river. International Agricultural, Biological & Life Science Conference, Edirne, Turkey, 2018.
7. Ahamad, K.U.; Raj, P.; Barbhuiya, N.H. Advances in Waste Management. *Springer Singapore* 2019. https://doi.org/10.1007/978–981–13–0215–2.
8. Oyebode, O.; Stretch, D. Neural network modeling of hydrological systems: A review of implementation techniques. *Nat. Resour. Model.* **2019**, *32*, 1–14. https://doi.org/10.1111/nrm.12189.
9. Hằng, H.T.M.; Hùng, N.T.; Dũng, N.V. Quản lý thống nhất và tổng hợp các nguồn thải gây ô nhiễm trên lưu vực hệ thống sông Đồng Nai. *Tạp chí Phát triển Khoa học và Công nghệ trẻ* **2006**, *9*, 5–17.
10. Gaurang, P.; Ganatra, A.; Kosta, Y.; Panchal, D. Behaviour Analysis of Multilayer Perceptronswith Multiple Hidden Neurons and Hidden Layers. *Int. J. Comput. Theory Eng*. **2011**, *3*, 332–337. https://doi.org/10.7763/IJCTE.2011.V3.328.
11. Le, V.T.; Nguyen, H.Q.; Loc, H.; Duyen, N.; Tran, D.; Duc, H.; Do, Q.H. A Multidisciplinary Approach for Evaluating Spatial and Temporal Variations in Water Quality. *Water* **2019**, *11*, 853.
12. Faris, H.; Aljarah, I.; Mirjalili, S. Chapter 28 – Evolving Radial Basis Function Networks Using Moth–Flame Optimizer. *in Handbook of Neural Computation, Pijush Samui, Sanjiban Sekhar, and Valentina E. Balas, Eds., ed: Academic Press*. **2017**, 537–550, ISBN: 978–0–12–811318–9.
13. Banadkookia, F.B.; Ehteram, M.; Panahic, F.; Sh. Sammend, S.; Othmane, F.B.; EL–Shafiee, A. Estimation of total dissolved solids (TDS) using new hybrid machine learning models. *J. Hydrol.* **2020**, *587*, 124989.
14. Ding, Y.R.; Cai, Y.J.; Sun, P.D.; Chen, B. The use of combined neural networks and genetic algorithms for prediction of river water quality. *J. Appl. Res. Technol.* **2014**, *12*, 493–499.
15. Elkiran, G.; Nourani, V.; Abba, S.I. Multi–step ahead modelling of river water quality parameters using ensemble artificial intelligence–based approach. *J. Hydrol*. **2019**, *577*, 123962.
16. Zhai, W.; Zhou, X.; Man, J.; Xu, Q.; Jiang, Q.; Yang, Z.; Jiang, L.; Gao, Z.; Yuan, Y.; Gao, W. Prediction of water quality based on artificial neural network with grey theory. *IOP Conf. Ser. Earth Environ. Sci*. **2019**, *295*, 042009.
17. Moriasi, D.N.; Gitau, M.W.; Pai, N.; Daggupati, P. Hydrologic and water quality Models performance measures and evaluation criteria. *Trans. ASABE Am. Soc. Agric. Biol Eng.* **2015**, 58, 1763–1785, https://doi.org/10.13031/trans.58.10715.

Simulation of Biochemical Oxygen Demand at the lower Sai Gon–Dong Nai Rivers using Artificial Neural Network models: Multilayer Perceptron (MLP) and Radial Basic Function (RBF)

Nguyen Thi Diem Thuy[1*], Pham Thi Thao Nhi[2], Doan Thi Truc Man[3], Dao Nguyen Khoi [4]

[1] Institute for Environment and Resources, Vietnam National University Ho Chi Minh city; nguyenthidiemthuyapag@gmail.com;
[2] Institute for Computational Science and Technology; ptthaonhi@gmail.com
[3] Southern Regional Hydrometeorological Center; trucmandoan@gmail.com
[4] Faculty of Environment, University of Science, Vietnam National University Ho Chi Minh city; dnkhoi@hcmus.edu.vn

Abstract: Biochemical Oxygen Demand is one of the most crucial water quality parameters to assess of water pollution of rivers. Nevertheless, BOD needs longer periods (5 days) to get results. The objective of this research is to build a computational model based on the artificial neural networks, including Multilayer Perceptron Network (MLP), and Radial Basis Function network (RBF) for simulating BOD_5 in the lower Sai Gon – Dong Nai rivers, and to evaluate the simulation efficiency between MLP and RBF. Seven different input combinations were constructed using Pearson correlation coefficients between each water quality parameter (COD, DO, TSS, Coliform, $P–PO_4^{3–}$, T, and $N–NH4^{+}$) and BOD_5. Five years (2013 to 2018) of monthly data from eight water quality monitoring stations within the study area were compiled, which were divided into two sub–sets (ratio 75:25) for model training and model testing. The results indicated that both the models satisfactorily simulated $BOD_{5,}$ but the RBF model with the combinations of variables numbered 5 (COD, DO, TSS, Coliform, $P–PO_4^{3–}$) demonstrated the best performance, values of Nash–Sutcliffe efficiency (NSE), coefficient of determination (R^2), and root mean square error (RMSE) were 0,848, 0,865, and 0,454, respectively. The results of this research are also the foundation for short–term prediction of BOD_5, as well as the simulation of the other water quality parameters in the area.

Keywords: Biochemical Oxygen Demand; Artificial Neural Network; MLP; RBF, The lower Sai Gon–Dong Nai rivers.